大语言模型在研究假设检验中替代人类的潜力*

姜红丙　高琳　向玉琳　张晨熙　李旭泽
（郑州大学管理学院 郑州 450001）

摘要：为探讨大语言模型替代人类学者进行研究假设检验的潜力，本研究选取 2023-2024 年管理学领域的 109 篇使用问卷调查法的论文为数据来源，提取出 646 个研究假设及相应的假设检验结果；然后，将每个研究假设中的所有概念及其定义输入 GPT-4 模型，输入提示语，让大语言模型判断这些概念间的关系，由 4 名作者分别解读模型输出的结果，并运用 Fleiss' Kappa 系数检验一致性；最后，运用 Kappa 系数分析大模型判断和论文研究假设、假设检验结果的一致性。研究说明大语言模型的判断与论文作者所提的研究假设、问卷调查检验的结果有着较高的一致性，显著高于随机水平，能够一定程度上替代人类检验研究假设。

关键词：大语言模型；假设检验；Kappa 系数
中图法分类号：C32；C39　　　**文献标识码：**A

The Potential of LLM to Replace Humans in Research Hypothesis Testing

Jiang Hongbing　Gao Lin　Xiang Yulin　Zhang Chenxi　Li XuZe
（School of Management, Zhengzhou University, Zhengzhou, China, 450001）

Abstract: In order to explore the potential of LLM replacing researchers to conduct research hypothesis testing, 109 research papers using survey method in the field of management from 2023 to 2024 were selected, and 646 research hypotheses and the results of the GPT-4 model, and then entering the prompt, the four authors interpreted the output results of the model respectively, and used Fleiss' Kappa coefficient to test the consistency; Finally, the consistency index Kappa coefficient is used to analyze the consistency of the results of the LLM judgment, the research hypotheses, and the tested results. The research suggests that the research hypothesis of the LLM test is highly consistent with the research hypotheses and the tested results, which is significantly higher than the random level, and can replace the human test research hypothesis to a certain extent.

Key words: LLM; hypothesis testing; Kappa coefficient

1. 研究背景

2022 年末 OpenAI 发布了聊天机器人 ChatGPT，使大语言模型（大模型）迅速普及，大模型成为近年来最热的话题。其核心是使用从互联网中提取的大量信息进行训练，包括文本和图像[1]。它能够在几乎所有主题中创建自然的、拟人化的专业文本，因此越来越多地被用作研究的辅助工具，社会科学研究也不例外[2]。

随着大模型的出现，研究者也逐渐探讨了其对社会科学研究的帮助，大模型可以在文献综述、问卷设计和数据分析等方面为社科研究提供助力。在文献综述方面，大模型能够智能化编写查询方法进行文献检索[3]，分类并总结现有文献，

* 本文系国家自然科学基金项目（项目编号：71801195；71961031）的研究成果之一。

生成层次结构化概览[4]，从而帮助学者快速了解某一领域的研究进展，更有效地进行文献综述[5]；在问卷设计方面，大模型可以从不同的公开数据源自动收集信息，根据研究需求智能生成问卷和调查设计[6]，并分类和整理数据[7]，节约研究者的时间成本；在数据处理与分析方面，大模型能够理解结构化和非结构化数据[8]，例如文本、访谈记录和社交媒体内容，并且快速执行数据清理[9]、文本分类[10]、主题分析[11]、情感分析[12]等复杂任务，显著减少了研究者的工作量。

此外，少量研究还探讨了大模型对于研究假设的提出和验证的助力。研究者们常常通过推理来提出研究假设，而大模型可以自动化地从大量原始网络语料库中提取信息[13]，识别可能的研究背景和启发性信息，进而生成研究假设。社会科学研究中提出的研究假设常常是用自然语言表达的陈述[14]，而大模型擅长从不同的语料中发现概念之间的联系，因此大模型对概念间关系的判断应该和研究者提出的研究假设有较大的一致性。Michael 使用 GPT-4 模型进行了尝试，说明了使用大模型来进行研究假设的生成和验证确实是可能实现的[15]，但是仍缺少证据证明大模型是否能够有效检验研究假设[16]。鉴于此，本研究将重点评估大模型在实证研究假设检验中的表现。据此，提出如下假设：

H1：大模型能够有效检验实证研究中的研究假设。

本文的核心研究问题是：大模型能否有效检验实证研究的研究假设？本研究将通过实证调查和数据分析，探讨大模型在社科领域中验证假设的潜力，以确定大模型在实证研究中的优势和不足。

2. 研究设计

本研究旨在评估大模型在实证研究中的应用效果，特别是其在假设检验中的表现。为此，本文收集管理学领域的实证分析文章，利用 GPT-4 模型对这些文章中提出的研究假设进行检验，并将模型的判断结果与文章中的实证分析结果进行对比。整体思路如图 1 所示：

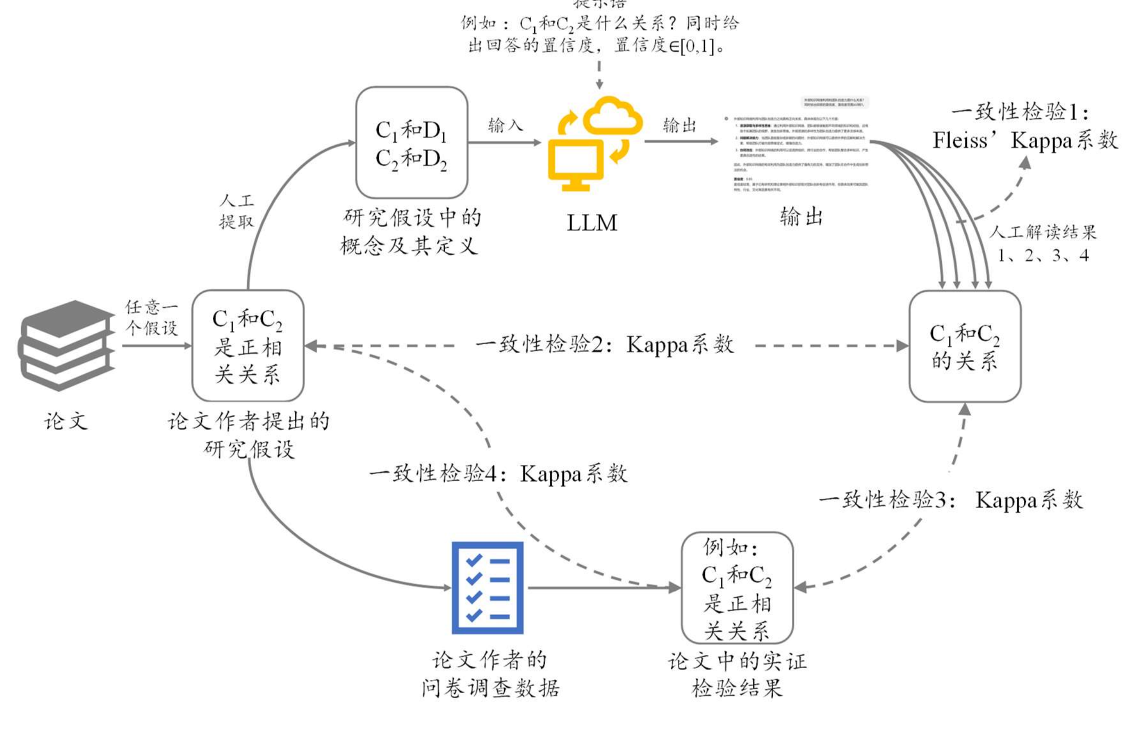


图 1 总体研究设计

2.1 数据来源

管理学研究在学术界具有重要影响力，特别是发表在高质量期刊上的实证研究，其结果往往被广泛引用和讨论。管理学作为一门综合性学科，涵盖多个研究主题，不同主题涉及多种类型的数据和研究方法，这使得大模型在不同情境下的应用效果可以得到全面评估。此外，在管理学研究中，假设研究检验是核心步骤之一，研究者常常通过检验研究假设来评价理论和实践结论。管理学领域高度依赖实证研究来检验研究假设，这也使得管理学成为探索大模型假设检验能力的理想领域。

本研究选择了《管理世界》《南开管理评论》《中国软科学》《管理评论》《科学学研究》和《管理学报》期刊，通过期刊官网收集 2023 年至 2024 年间发表在以上期刊上使用了问卷调查法的论文，作为实证分析的数据来源。选取以上论文作为数据来源有以下三个原因：

①本研究选择的都是经过严格筛选的高质量学术期刊，这些期刊发表的论文质量高，研究方法较为严谨，结果可靠性高。此外，这些期刊涵盖了广泛的管理学领域研究问题，包括但不限于企业管理、市场营销、组织行为和人力资源管理等。这保证了样本的多样性和研究结果的普遍性。

②选择 2023 年至 2024 年间发表的论文，可以确保所使用的数据和研究结果是相对较新的。使用这些数据可以更好地反映大模型的当前应用效果，这对于验证大模型在当前数据环境和研究趋势下的表现具有重要意义。

③问卷调查法在管理学研究中被广泛应用，适用于多种研究问题和情境。其通常使用结构化的问卷，直接从研究对象处获得数据，通常数据的真实性和可靠性较高。此外，使用了问卷调查法的论文大多都明确提出了研究假设并进行了实证检验，并且文中概念定义往往较为清晰明确，这方便本研究提取数据。

2.2 数据收集

首先，手工从每一篇论文中提取所有研究假设，再从论文中提取研究假设中的重点概念和其理论定义。此外，还需要提取论文作者通过问卷调查法得出的实证检验结果（即假设被接受或拒绝）。当研究假设为两个概念 C_1 和 C_2 之间的关系时，将概念 C_1 及其定义 D_1 输入到 GPT-4 模型中，再将概念 C_2 及其定义 D_2 也输入到 GPT-4 模型中，让其先理解相关概念和研究假设的背景。接着输入提示语：C_1 和 C_2 是什么关系？同时给出回答的置信度，置信度 $\in[0,1]$。模型输出置信度较低（<0.6）的回答则不予采用，让其重新回答。得到可用的输出后，本研究安排 4 个作者分别解读 GPT-4 模型给出的回答。同理，当研究假设为三个或者多个概念之间的关系，也是先将每个概念及定义作为输入，再输入提示语询问这些概念之间的关系及回答的置信度，最后分别人工解读结果。完成之后，将每个结果转化成对假设的判断，并记录下最终的判断结果（成立√或不成立×）。

以 2023 年《南开管理评论》上刊登的《内外兼修：多重网络交互影响下的自组织团队创造力研究》一文为例，该文探讨了在数字技术深入应用背景下，自组织团队内部网络去中心化和外部知识网络利用如何交互影响团队的创造力，其中一个研究假设是“外部知识网络利用能够正向影响团队创造力”。本研究将“外部知识网络利用”和“团队创造力”的概念提取出来，输入 GPT-4 模型中，如图 2 所示。

外部知识网络利用的概念：外部知识网络利用是指团队从组织内团队外与组织外的成员中获取相关知识和资源的能力。
团队创造力的概念：团队创造力在本质上具有协同性，是团队成员互动和相互影响的产物。

外部知识网络利用的概念强调团队从外部获取知识和资源的能力，这些知识和资源可以来自组织内部的其他团队、组织外部的合作伙伴、行业专家或客户等。它反映了团队如何通过外部的合作与交流来丰富自身的知识储备，以更好地应对复杂的问题或推动创新。

团队创造力的概念则聚焦于团队成员之间的协同合作。团队创造力不是单一成员的产物，而是通过成员间的互动、思想碰撞和相互影响所形成的整体创造性成果。有效的团队创造力依赖于团队成员的多样化思维、共享的目标和良好的沟通机制。

两者结合时，可以说外部知识网络的利用有助于激发团队的创造力，拓展了团队的思维广度，为创造性的合作提供了更多的输入和支持。

图 2 研究假设中的概念及定义输入 GPT-4 模型

再输入提示语：外部知识网络利用和团队创造力是什么关系？同时给出回答的置信度，置信度∈[0,1]。得到的输出如图 3 所示，人工解读该回答得到的概念之间的关系是外部知识网络利用能够正向影响团队创造力，即模型支持论文作者所提出的研究假设。

外部知识网络利用和团队创造力是什么关系？
同时给出回答的置信度，置信度∈[0,1]。

外部知识网络利用和团队创造力之间的关系是相辅相成的。外部知识网络利用可以增强团队创造力，具体表现在以下几个方面：

1. **丰富知识来源**：团队通过从外部获取新知识和多元的资源，能够拓展思维边界，打破常规的思维模式，提供更多创新的可能性。外部知识的引入往往带来新的视角或方法，有助于团队解决复杂问题，推动创新。
2. **增强协同合作**：外部知识网络不仅限于资源的获取，还包括与外部合作伙伴的互动与交流，这为团队内部成员提供了更多的合作机会和创意碰撞的场所，有助于提升团队的创造性成果。
3. **提高创新能力**：通过有效地利用外部知识，团队能够更好地理解行业趋势、技术发展或市场需求，这些信息可以帮助团队设计更具前瞻性和市场竞争力的创意或产品。

综合来看，外部知识网络利用为团队创造力的提升提供了外部支持和资源，而团队创造力则是通过有效利用这些外部资源来实现创新的过程。

置信度：0.85
这个结论基于对已有研究和理论的分析，大部分学者认为外部知识网络的有效利用能显著促进团队创造力，尽管不同情境下两者的关系可能有所差异。

图 3 GPT-4 模型输出结果

2.3 数据分析

本研究将分四个检验步骤来确定大模型在研究假设检验中的有效性，如图 4

所示。首先检验4名作者人工解读大模型输出结果的一致性（检验1）。如果一致性>0.6（中度一致性水平），则说明4名作者解读的结果一致性较高，结果较为可靠，可以进行下一步检验；一致性≤0.6，说明4名作者解读的结果主观性较高，需要重新解读。下一步先检验大模型判断结果和论文研究假设的一致性（检验2），接着检验大模型判断和假设检验结果的一致性（检验3），看两者的结果是否高于随机一致性，如果都高于则将两者一致性结果进行对比。最后，检验论文研究假设和假设检验结果的一致性（检验4），并将结果与检验2进行对比。

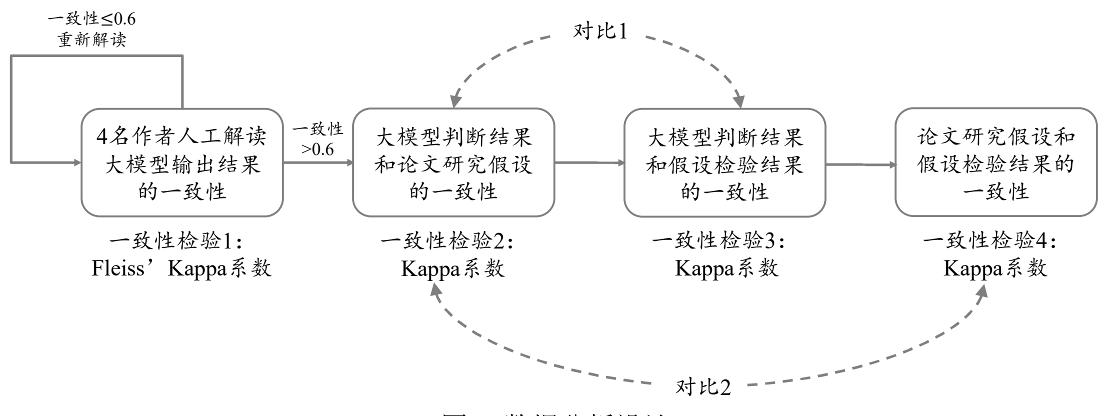


图4 数据分析设计

（1）检验1：检验4名作者人工解读大模型输出结果的一致性

本研究采用Fleiss’Kappa系数来衡量4名作者人工解读大模型输出结果的一致性。Fleiss’Kappa是一种扩展的Kappa系数，用于评估三个及以上评分者对多个项目进行评分分类时的一致性。Fleiss’Kappa系数的解释通常遵循以下标准：0<k≤0.40，一致性较差；0.40<k≤0.60，中度一致；0.60<k≤0.80，一致性较高；k>0.80，一致性极好。

统计收集到的4名作者人工解读结果1、2、3、4，构建数据矩阵，根据数据矩阵计算Fleiss’Kappa系数，比对4名作者解读结果的一致性，计算步骤如下。

①计算出观察到的一致性P_0

先计算每个研究假设的一致性P_i：

$$P_i = \frac{\Sigma_m n_{im}(n_{im} - 1)}{N \times (N - 1)}$$

其中n_{im}是第i个研究假设被判断为m（即√或×）的次数，N为第i个研究假设判断结果的总数量。计算完所有P_i后，取平均值作为总体一致性P_0：

$$P_0 = \frac{\Sigma P_i}{n}$$

②计算偶然一致性$\mathrm{P_e}$：

$$P_e = \Sigma P_m^2$$

其中P_m是研究假设被判断为m（即√或×）的概率。

③计算Fleiss’Kappa系数：

$$k = \frac{P_0 - P_e}{1 - P_e}$$

计算出k值后，根据Fleiss’Kappa系数的解释标准来分析结果，评估4名作者解读结果的一致性。如果一致性高于0.6，则说明本研究人工解读结果可靠

性较高，可代表大模型的判断结果进行下述检验。

（2）检验 2、3：检验大模型判断结果和论文研究假设、假设检验结果的一致性

本研究采用 Kappa 系数来衡量大模型判断结果和论文研究假设、假设检验结果的一致性。这是一种衡量两个评估者之间一致性的指标，它考虑了偶然一致的情况，因此比简单的百分比一致性更为可靠。Kappa 系数的值范围从-1 到 1，其中 1 表示完全一致，0 表示没有超出随机预期的一致性，负值表示一致性低于随机预期。

统计收集到的论文作者提出的研究假设和假设检验的结果，再统计 4 名作者人工解读的 GPT-4 模型生成的研究假设检验结果，将上述情况进一步比对，看其中有多少个由论文作者提出的研究假设与 GPT-4 模型生成结果（人工解读结果 1、2、3、4）一致，又有多少个经实证检验过的结果与 GPT-4 模型生成结果（人工解读结果 1、2、3、4）一致。据此分别建立 2×2 列联表，如表 1 所示，根据列联表计算出观察到的一致性比例P_0和期望的一致性比例P_e（本文取随机概率 0.5），再使用 Kappa 系数计算出一致性比例，公式如下。

表 1 2×2 列联表

判断 1 VS 判断 2	√	×	合计
√	a	b	a+b
×	c	d	c+d
合计	a+c	b+d	N

$$P_0 = \frac{\mathrm{a + d}}{\mathrm{N}}$$

$$\mathrm{Kappa} = \frac{\mathrm{P_0 - P_e}}{\mathrm{1 - P_e}}$$

计算出人工解读结果 1、2、3、4 和论文研究假设、假设检验结果的 Kappa 值后，取平均值作为 GPT-4 模型判断结果和论文研究假设、假设检验结果的一致性。再根据 Kappa 系数的解释标准来分析结果，评估大模型判断结果和论文研究假设、假设检验结果之间的一致性。如果两者一致性都高于随机水平，则将两者进行对比分析，看 GPT-4 模型判断结果是与论文研究假设的一致性更高，还是与假设检验结果的一致性更高。

（3）检验 4：检验论文研究假设和假设检验结果的一致性

本研究采用 Kappa 系数来衡量论文研究假设和假设检验结果的一致性。统计收集到的论文作者提出的研究假设和假设检验的结果，看有多少论文作者提出的研究假设与假设检验结果一致，据此建立 2×2 列联表。根据列联表计算出观察到的一致性比例P_0和期望的一致性比例P_e（本文取随机概率 0.5），再使用 Kappa 系数计算出一致性比例，公式同检验 2、3。

计算出 Kappa 值后，将该一致性结果与检验 2 结果比对，看大模型与传统检验方法的差距，以进一步分析大模型在研究假设检验中的优势和不足

3. 实证结果

3.1 数据描述

在本研究中，共分析了 2023 年至 2024 年间发表在《管理世界》《南开管理

评论》《中国软科学》《管理评论》《科学学研究》和《管理学报》上的 109 篇实证分析论文，所有选取的论文均采用了问卷调查法进行数据收集。这些论文涵盖了管理学领域多个研究主题，包括组织行为、人力资源管理、创新管理、市场营销、企业伦理与社会责任、战略管理和财务管理等，具体如表 2 所示。

表 2 论文主题分布

主题	主要关注问题	数量
组织行为	团队协作、领导力、团队沟通、集体决策、团队绩效	27
人力资源管理	人才管理、绩效考核、员工培训、激励机制、职场满意度	25
创新管理	技术突破、研发投入、创新扩散、专利产出、技术采纳	23
市场营销	购买意愿、品牌忠诚、顾客满意度、消费者决策过程、市场细分	17
企业伦理与社会责任	可持续发展、企业公民、伦理决策、环境责任、社会影响	9
战略管理	战略规划、竞争优势、市场定位、资源配置、战略执行	5
财务管理	财务分析、资本结构、风险管理、投资决策、财务报表	3

进一步分析这些论文的描述性结果，本研究统计了每篇论文提出的研究假设数量其验证通过率。结果显示，总研究假设数量为 646 个，其中，有 606 个研究假设得到了经验支持，经验支持率高达 93.8%；剩余的 40 个研究假设未能得到经验支持，未得到经验支持的概率为 6.2%。

此外，本研究对人工解读大模型的检验结果进行统计与分析，观察到的一致性比例如表 3 所示。在 646 个研究假设中，由人工解读大模型输出的结果 1，有 523 个研究假设成立结果与论文研究假设一致，一致率达到了 81.0%；511 个研究假设的大模型判断结果与论文检验结果相同，一致率为 79.1%。由人工解读大模型输出的结果 2，有 530 个研究假设成立结果与论文研究假设一致，一致率为 82.0%；520 个研究假设的大模型判断结果与论文检验结果相同，一致率为 80.5%。由人工解读大模型输出的结果 3 中，有 533 个研究假设成立结果与论文研究假设一致，一致率为 82.5%；517 个研究假设的大模型判断结果与论文检验结果相同，一致率为 80.0%。由人工解读大模型输出的结果 4 中，有 569 个研究假设成立结果与论文研究假设一致，一致率高达 88.1%；557 个研究假设的大模型判断结果与论文检验结果相同，一致率高达 86.2%。这一现象表明，从观察到的一致性比例来看，大模型生成的文本与论文研究假设和论文检验结果存在比较高的一致性。

表 3 结果一致性

结果一致性		一致	不一致
大模型判断结果与论文研究假设比较	人工解读结果 1	81.0%	19.0%
	人工解读结果 2	82.0%	18.0%
	人工解读结果 3	82.5%	17.5%
	人工解读结果 4	88.1%	11.9%
大模型判断与假设检验结果比较	人工解读结果 1	79.1%	20.9%
	人工解读结果 2	80.5%	19.5%
	人工解读结果 3	80.0%	20.0%
	人工解读结果 4	86.2%	13.8%

通过分析论文的研究主题、研究假设经验支持率和观察到的大模型判断结果一致性，本研究初步了解管理学领域实证研究的全面概况和大模型在检验实证研

究假设中的表现和潜力。然而，大模型在实证研究假设检验中的具体表现如何，还需要进行假设检验。

3.2 假设检验

本研究在假设检验部分将针对 H1 假设，即“大模型能够有效检验实证研究中的假设”进行分析，评估大模型在假设检验中的应用效果。

（1）检验 1：检验 4 名作者人工解读大模型输出结果的一致性

首先，建立数据矩阵，记录人工解读结果 1、2、3、4 的一致性和不一致性情况，其中每一行代表一个研究假设，每一列代表对研究假设判断的结果个数。例如，第一行中的数字“4”代表着 4 名作者都认为大模型判断研究假设 H1 正确，数字“0”代表着没有作者认为大模型判断研究假设 H1 错误。本研究以文章《母国制度情境下逆向知识转移的长期创新效应》的数据矩阵为例，如表 4 所示，为 109 篇文章中的 646 个研究假设检验结果建立数据矩阵。

表 4 人工解读结果 1、2、3、4 一致性数据矩阵

研究假设	√	×
H1.逆向知识转移对母公司的长期创新绩效具有积极影响。	4	0
H2.双元平衡在逆向知识转移与长期创新绩效之间发挥了中介作用。	0	4
H3.双元联合在逆向知识转移与长期创新绩效之间发挥了中介作用。	1	3
H4.母国制度情境正向调节逆向知识转移与双元平衡之间的关系。	3	1
H5.母国制度情境正向调节逆向知识转移与双元联合之间的关系。	2	2
H6.母国制度情境正向调节双元平衡在逆向知识转移与长期创新绩效之间的中介作用。	0	4
H7.母国制度情境正向调节双元联合在逆向知识转移与长期创新绩效之间的中介作用。	1	3

①计算出观察到的一致性P_0

先计算每个研究假设的一致性P_i，如对上例研究假设 H1：

$$P_1 = \frac{4 \times (4-1) + 0 \times (0-1)}{4 \times (4-1)} = 1$$

对其他研究假设重复计算P_i，取平均值作为总体一致性P_0：

$$P_0 = \frac{\Sigma P_i}{n} = \frac{594.83}{646} = 0.92$$

②计算偶然一致性$\mathrm{P_e}$：

$$P_e = \Sigma P_m^2 = \sqrt{}\text{的比率}^2 + \times\text{ 的比率}^2 = \frac{2155^2 + 429^2}{2584^2} = 0.72$$

③计算 Fleiss’Kappa 系数：

$$k = \frac{P_0 - P_e}{1 - P_e} = \frac{0.92 - 0.71}{1 - 0.71} = 0.71$$

本研究计算出的 k 值为 0.71，大于 0.6，这意味着 4 名作者人工解读的结果有着较高的一致性。由于 4 名作者是独立地进行解读，说明大模型的输出具有一

定的辨识度，不会产生过多的错误解读。

（2）检验 2：检验大模型判断结果和论文研究假设的一致性

首先，建立四个 2x2 列联表，分别记录大模型判断结果（人工解读结果 1、2、3、4）和论文研究假设的一致性和不一致性情况，如表 5、6、7、8 所示。其中，行表示论文研究假设（默认为√），列表示大模型判断结果（√或×），单元格的值表示对应判断组合的研究假设数量。例如，表 5 中第一行的数字“523”代表着有 523 个论文研究假设是人工解读结果 1 判断为正确的，数字“123”代表着有 123 个论文研究假设是人工解读结果 1 判断为错误的。

表 5 人工解读结果 1 VS 论文研究假设

人工解读结果 1	√	×	总计
√	523	123	646
×	0	0	0
总计	523	123	646

表 6 人工解读结果 2 VS 论文研究假设

人工解读结果 2	√	×	总计
√	530	116	646
×	0	0	0
总计	530	116	646

表 7 人工解读结果 3 VS 论文研究假设

人工解读结果 3	√	×	总计
√	533	113	646
×	0	0	0
总计	533	113	646

表 8 人工解读结果 4 VS 论文研究假设

人工解读结果 4	√	×	总计
√	569	77	646
×	0	0	0
总计	569	77	646

接着由列联表计算出观察到的一致性比例P_0，期望的一致性比例$\mathrm{P_e}$（本文取随机概率 0.5），代入公式计算 Kappa 值。

对于人工解读结果 1：

$$P_0 = \frac{523 + 0}{646} = 0.81，Kappa = \frac{P_0 - P_e}{1 - P_e} = \frac{0.81 - 0.5}{1 - 0.5} = 0.62$$

对于人工解读结果 2：

$$P_0 = \frac{530 + 0}{646} = 0.82，Kappa = \frac{P_0 - P_e}{1 - P_e} = \frac{0.82 - 0.5}{1 - 0.5} = 0.64$$

对于人工解读结果 3：

$$P_0 = \frac{533 + 0}{646} = 0.83，Kappa = \frac{P_0 - P_e}{1 - P_e} = \frac{0.83 - 0.5}{1 - 0.5} = 0.65$$

对于人工解读结果 4：

$$P_0 = \frac{569 + 0}{646} = 0.88，Kappa = \frac{P_0 - P_e}{1 - P_e} = \frac{0.88 - 0.5}{1 - 0.5} = 0.76$$

对人工解读结果 1、2、3、4 的 Kappa 值取平均数：

$$\overline{Kappa} = \frac{\Sigma Kappa}{n} = \frac{0.62 + 0.64 + 0.65 + 0.76}{4} = 0.67$$

由上述计算结果可知，大模型判断结果和论文研究假设的一致性为 0.67，在

0.6 和 0.8 之间，拥有较高的一致性，这意味着大模型判断结果和论文研究假设的一致性很高，远超过随机水平，也说明了大模型的判断结果具有较高的可靠性。

（3）检验 3：检验大模型判断和假设检验结果的一致性

首先，建立四个 2x2 列联表，分别记录大模型判断结果（人工解读结果 1、2、3、4）和假设检验结果的一致性和不一致性情况，如表 9、10、11、12 所示。其中，行表示假设检验结果（√或×），列表示大模型判断结果（√或×），单元格的值表示对应判断组合的研究假设数量。例如，表 9 中第一行的数字“497”代表着有 497 个研究假设是假设检验结果正确且人工解读结果 1 判断为正确的，数字“109”代表着有 109 个研究假设是假设检验结果正确但人工解读结果 1 判断为错误的。

表 9 人工解读结果 1 VS 假设检验结果

人工解读结果 1	√	×	总计
√	497	109	606
×	26	14	40
总计	523	123	646

表 10 人工解读结果 2 VS 假设检验结果

人工解读结果 2	√	×	总计
√	505	101	606
×	25	15	40
总计	530	116	646

表 11 人工解读结果 3 VS 假设检验结果

人工解读结果 3	√	×	总计
√	505	101	606
×	28	12	40
总计	533	113	646

表 12 人工解读结果 4 VS 假设检验结果

人工解读结果 4	√	×	总计
√	543	63	606
×	26	14	40
总计	569	77	646

接着由列联表计算出观察到的一致性比例P_0，期望的一致性比例$\mathrm{P_e}$（本文取随机概率 0.5），代入公式计算 Kappa 值。

对于人工解读结果 1：

$$P_0 = \frac{497 + 14}{646} = 0.79，Kappa = \frac{P_0 - P_e}{1 - P_e} = \frac{0.79 - 0.5}{1 - 0.5} = 0.58$$

对于人工解读结果 2：

$$P_0 = \frac{505 + 15}{646} = 0.81，Kappa = \frac{P_0 - P_e}{1 - P_e} = \frac{0.81 - 0.5}{1 - 0.5} = 0.61$$

对于人工解读结果 3：

$$P_0 = \frac{505 + 12}{646} = 0.80，Kappa = \frac{P_0 - P_e}{1 - P_e} = \frac{0.80 - 0.5}{1 - 0.5} = 0.60$$

对于人工解读结果 4：

$$P_0 = \frac{543 + 14}{646} = 0.86，Kappa = \frac{P_0 - P_e}{1 - P_e} = \frac{0.86 - 0.5}{1 - 0.5} = 0.72$$

对人工解读结果 1、2、3、4 的 Kappa 值取平均数：

$$\overline{Kappa} = \frac{\Sigma Kappa}{n} = \frac{0.58 + 0.61 + 0.60 + 0.72}{4} = 0.63$$

上述计算出的$\overline{Kappa}$值为 0.63，在 0.6 和 0.8 之间，这意味着大模型判断结果和假设检验结果的一致性较高。

（4）检验 4：检验论文研究假设和假设检验结果的一致性

首先，建立一个 2x2 列联表，记录论文研究假设和假设检验结果的一致性和不一致性情况，如表 13 所示。其中，行表示论文研究假设（默认为√），列表示假设检验结果（√或×），单元格的值表示对应判断组合的研究假设数量。表 13 中第一行的数字"606"代表着有 606 个论文研究假设是假设检验结果正确的，数字"40"代表着有 40 个论文研究假设是假设检验结果错误的。

表 13 论文研究假设 Vs 假设检验结果

论文研究假设 VS 假设检验结果	√	×	总计
√	606	40	646
×	0	0	0
总计	606	40	646

接着由列联表计算出观察到的一致性比例P_0，期望的一致性比例P_e（本文取随机概率 0.5），代入公式计算 Kappa 值。

$$P_0 = \frac{606+0}{646} = 0.94，Kappa = \frac{P_0 - P_e}{1 - P_e} = \frac{0.94 - 0.5}{1 - 0.5} = 0.88$$

上述计算出的$Kappa$值为 0.88，在 0.8 和 1 之间，即论文研究假设和假设检验结果之间有着高度一致性。

以上这些检验结果说明了大模型判断结果与研究假设、假设检验结果的一致性很高，远超过随机水平，证实了大模型在检验实证研究假设中的有效性，即 H1 得到支持。这说明大模型不仅能够正确理解研究假设和相关数据，还能够在复杂的分析场景中做出较为准确的预测。这一表现展示了其在学术研究中，尤其是在处理大量数据和评估研究假设时的潜在应用价值。从这个角度来看，大模型提供了一种可靠的工具，来提高研究假设检验的效率和准确性。尽管大模型并非在所有情况下都能给出完全正确的判断，但其成功率接近问卷调查法的水平，说明了其在科研应用中的潜力。

接下来，为了更深入地探究大模型在检验研究假设中的优势和不足，本研究将根据上述检验结果进行对比。

（5）检验 2 与检验 3 对比

检验 2 计算出的一致性为 0.67，高于检验 3 的一致性 0.63，也就是说大模型判断结果与研究假设的一致性要略高于与假设检验结果的一致性，这有可能是由于大模型的输出主要来自文本语料的训练，而非问卷调查数据。

（6）检验 2 与检验 4 对比

检验 4 计算出的一致性为 0.88，高于检验 2 的 0.67，即论文研究假设与假设检验结果的一致性要高于与大模型判断结果的一致性。这表明虽然大模型判断与假设检验结果在很大程度上是一致的，也就是说大模型在假设检验中的表现与问卷调查法接近，但是还有一定的距离。

3.3 结果分析

（1）大模型的优势

大模型在检验研究假设时，展现出了快速且优秀的检验能力，具有良好的应

用潜力。传统的假设检验方法通常需要研究人员手动进行数据分析、模型构建和结果验证等，这些过程往往耗费大量时间和人力资源。相比于传统的检验方法，大模型几乎不消耗人力物力以及时间，轻松就能得到很大程度上有用的结果。从这个角度来说，使用大模型帮助学者预测或是检验假设，减轻了人工分析的负担，节约了大量的研究时间和资源，给实证研究带来了很大的助力。

（2）大模型的局限性

虽然大模型在假设检验中的表现与实证检验结果拥有较大的一致性，但与问卷调查法检验假设的验证高通过率还有一定的距离。这也在一定程度上反映了大模型在大多数场景下表现良好，但是还缺乏处理复杂或涉及深层理论推理的假设的能力，仍有改进空间。复杂的假设往往需要更多的背景信息和细致的变量控制，这些可能超出了大模型的推理能力。在这些情况下，传统的检验方式仍然不可或缺。因此，为了提升大模型的检验效果，未来应进一步探索结合传统统计方法与大模型的混合分析方法。

4. 讨论与展望

本研究的理论贡献在于：本研究为大模型在实证研究中的应用提供了新视角，填补了大模型在假设检验领域的研究空白。本研究将大模型用于实证研究的假设检验，验证其在假设检验中的有效性，进一步扩展了大模型的应用范围。许多研究已指出大模型在数据处理和文本分析上的优势，尤其在 NLP 领域，大模型的表现已被广泛认可。然而，与现有研究相比，本研究也揭示了一些新的发现，专注于大模型在假设检验这一任务中的应用效果。这是对现有理论的扩展，表明大模型不仅可以用于分析和处理数据，还可以有效进行假设检验。这一应用拓展了大模型在学术研究中的使用范围，为未来大模型在其他研究领域的进一步应用提供了理论基础。

尽管本研究取得了一些有意思的发现，但仍存在几点局限：①本研究仅选择了 2023 年至 2024 年间发表的文章作为数据来源，数量和时间范围较为有限。未来研究可以通过扩大数据来源范围和时间跨度，获得更具代表性的结果。②本研究主要关注管理学领域的实证研究，尚未涵盖其他领域的文章。未来研究可以涵盖更多不同领域的研究，探索大模型在其他学科中的表现。③未来可以结合传统的人工分析方法和大模型，设计混合分析流程，从而增强大模型在复杂假设检验中的表现。

总之，本研究通过对大模型在实证研究假设检验中的表现进行评估，明确了其在假设检验中的有效性。大模型在假设检验中表现出较高的成功率，虽然未达到问卷调查法检验的水平，但其能够在较短时间内提供高效的假设检验，节省时间和人力成本的优势非常显著。

参考文献

[1] Liu Y, Han T, Ma S, et al. Summary of ChatGPT-related research and perspective towards the future of large language models[J]. Meta-Radiology, 2023, 1(2): 100017.

[2] Bail C A. Can generative AI improve social science?[J]. Proceedings of the National Academy of Sciences, 2024, 121(21): e2314021121.

[3] Wang S, Scells H, Koopman B, et al. Can ChatGPT write a good boolean query for systematic review literature search?[C]. New York, NY, USA: Association for Computing Machinery, 2023: 1426-1436.

[4] Hsu C C, Bransom E, Sparks J, et al. Chime: LLM-assisted hierarchical organization of scientific studies for literature review support[J]. arXiv preprint arXiv:2407.16148, 2024.

[5] Antu S A, Chen H, Richards C K. Using LLM (large language model) to improve efficiency in literature review for undergraduate research[C]. Tokyo, Japan: AIED, 2023: 8-16.

[6] Zou Z, Mubin O, Alnajjar F, et al. A pilot study of measuring emotional response and perception of LLM-generated questionnaire and human-generated questionnaires[J]. Scientific Reports, 2024, 14(1): 2781.

[7] ZiHang C, QianMin S, GaoYi C, et al. Application scheme of clinical trial questionnaire pre recruitment integrating LLM and knowledge graph[J]. Available at SSRN 4713177, 2024.

[8] Sui Y, Zhou M, Zhou M, et al. Table meets LLM: Can large language models understand structured table data? A benchmark and empirical study[C].New York, NY, USA: Association for Computing Machinery, 2024: 645-654.

[9] Lee C T. Can an LLM find its way around a spreadsheet?[D]. Virginia Tech, 2024.

[10] Martinez C, Miller A, Garcia J, et al. The advancements and applications of AIGC large models[J]. DOI:10.13140/RG.2.2.14596.91524, 2023.

[11] Chew R, Bollenbacher J, Wenger M, et al. LLM-assisted content analysis: Using large language models to support deductive coding[J]. arXiv preprint arXiv:2306.14924, 2023.

[12] Bhat R,Jain B. Stock price trend prediction using emotion analysis of financial headlines with distilled LLM model[C]. New York, NY, USA: Association for Computing Machinery, 2024: 67–73.

[13] Yang Z,Du X,Li J, et al. Large language models for automated open-domain scientific hypotheses discovery[J]. arXiv:2309.02726, 2024.

[14] Jaccard J,Jacoby J. Theory construction and Model-Building skills: A practical guide for social scientists[M]. Guilford Publications, 2020.

[15] King M. Can GPT-4 formulate and test a novel hypothesis? Yes and no[J]. TechRxiv, 2023.

[16] Xu R,Sun Y,Ren M,et al. AI for social science and social science of AI: A survey[J]. Information Processing & Management, 2024, 61(3): 103665.